A Two-Level Interpretation of Modality in Human-Robot Dialogue

Lucia Donatelli[1], Kenneth Lai[2], James Pustejovsky[2]
[1]Department of Language Science and Technology
Saarland University, Germany
[2]Department of Computer Science
Brandeis University, USA
donatelli@coli.uni-saarland.de
{klai12,jamesp}@brandeis.edu

Abstract

We analyze the use and interpretation of modal expressions in a corpus of situated human-robot dialogue and ask how to effectively represent these expressions for automatic learning and dynamic interpretation in context. We present a two-level annotation scheme for modality that captures both content and intent, integrating a logic-based, semantic representation and a task-oriented, pragmatic representation that maps to our robot's capabilities. Data from our annotation task reveals that the interpretation of modal expressions in human-robot dialogue is quite diverse, yet highly constrained by the physical environment and asymmetrical speaker/addressee relationship. We sketch a formal model of human-robot common ground in which modality can be grounded and dynamically interpreted relative to speaker role, temporal constraints, and physical environment.

1 Introduction

The interpretation of modal expressions is essential to meaningful human-robot dialogue: the ability to convey information about objects and events that are displaced in time, space, and actuality allows the human and robot to align their environmental perceptions and successfully collaborate (Liu and Chai, 2015). As an example, if a robot is sent to a remote location on a search and navigation mission, modally interpreted expressions such as "*Tell me what you see*" (uttered by the human) and "*I can't see because of smoke*" (uttered by the robot) are vital to information exchange. Similarly, a robot that has abilities to navigate obstacles (for example, by jumping or using LIDAR) can inform the human of this.

The learning of modal expressions for automatic understanding and use nevertheless presents a conversational paradox: while these expressions serve to communicate and align world knowledge, there is no obvious manner to ground them in the shared environment. Whereas objects and actions can be pointed to or modeled for grounded learning, modal expressions are grounded in the linguistic signal itself. Nevertheless, a basic understanding of modal meaning would allow non-human agents to reason about the possible uses of objects and better assess how certain actions and behaviors impact the task at hand.

In this paper, we document the range and nature of modally interpreted expressions used in human-robot dialogue with the goal to make the interpretation of such expressions easily automated in the future. We hypothesize that certain readings and scope preferences for modal operators are more salient in human-robot dialogue because of the unique makeup of the common ground (Poesio, 1993). We provide a mapping from formal semantic theories of modality related to participant beliefs and updates of the common ground (Portner, 2009), to a practical model of speech acts that translates into robot action for search and navigation task-oriented dialogue and an automated NLU and NLG system (Bonial et al., 2020). This mapping is formalized in an annotation scheme in which the use of modal expressions is mapped to their effect in dialogue, providing a model for the robot to learn the meaning of modal expressions (Chai et al., 2018). Our annotation task reveals surprisingly high inter-annotator agreement for a complex scheme; results indicate that our data is highly repetitive in the natural language used,

Proceedings of the 28th International Conference on Computational Linguistics, pages 4222–4238
Barcelona, Spain (Online), December 8-13, 2020

and yet the interpretation of modal expressions is quite diverse and worth investigating further to foster effective human-robot communication in situated, task-oriented settings.

The paper is structured as follows. In Section 2 we motivate our annotation of modality and introduce the SCOUT corpus; we then situate formal semantic theories of modality in the context of human-robot dialogue. We describe our annotation scheme in Section 3, which covers both the type of modality used in an expression, and the speech act the expression conveys. We describe our results in Section 4, discussing implications for modal interpretation in human-robot dialogue and some linguistic issues that arose during the annotation process. In Section 5 we consider the implications of our results for a theory of modality and common ground in human-robot dialogue, before concluding in Section 6.

2 Background and Related Work

2.1 Annotating Modality

Though there is little previous work on annotating modality in dialogue, several annotation schemes exist for annotating modality in text. These annotation schemes often address modality and negation together as extra-propositional aspects of meaning, focusing on the tasks of detecting key linguistic markers and mapping their scope (Saurı et al., 2006; Morante and Daelemans, 2012). These tasks can then be leveraged to identify and analyze related concepts such as subjectivity, hedging, evidentiality, uncertainty, committed belief, and factuality (Morante and Sporleder, 2012). Automatic tagging of modality and negation and detection of related concepts has received little, though promising, attention (Baker et al., 2012; Prabhakaran et al., 2012; Marasović and Frank, 2016). The detection of the related concept of hedging (and its scope) was the focus of the CoNLL 2010 Shared Task (Farkas et al., 2010).

In the context of human-robot dialogue, modality and related concepts provide the basis for assessing speaker beliefs, commitments, and attitudes, thereby fostering understanding and coherent interaction. For example, the manner in which a speaker employs modal information can be used to assess trustworthiness (Su et al., 2010); this is important both for the human to trust the robot and work collaboratively, and for the robot to assess whether or not it should accept human instruction. Additionally, modal information allows both dialogue participants to assess the factuality of events and propositions (Saurí and Pustejovsky, 2009; Prabhakaran et al., 2015). Notably this is a complex process that requires the understanding of both fine-grained lexical semantics (e.g. a question to the robot, *"Can you prevent fire?"*) and the interaction of scopal operators (e.g. a robot assertion, *"I probably cannot fit there."*). Our work is one step towards aligning participant perceptions of respective environments and both discourse and real-world events.

2.2 Human-Robot Dialogue

The data we annotate comes from the Situated Corpus of Understanding Transactions (SCOUT), a collection of dialogues from the robot navigation domain.[1] SCOUT was created to explore the natural diversity of communication strategies in situated human-robot dialogue (Marge et al., 2016; Marge et al., 2017). Data collection efforts leveraged "Wizard-of-Oz" experiment design (Riek, 2012), in which participants directed what they believed to be an autonomous robot to complete search and navigation tasks. The domain testbed for this data was collaborative exploration in a low-bandwidth environment, mimicking the conditions of a reconnaissance or search-and-navigation operation. For data collection, two "wizard" experimenters controlled the robot's dialogue processing and robot navigation capabilities behind the scenes. This design permitted participants to instruct the robot without imposing artificial restrictions on the language used. As more data was collected, increasing levels of automated dialogue processing were introduced (Lukin et al., 2018a). We discuss the impact of further design details in Sections 4 and 5.

Table 1 shows an example SCOUT interaction. The dialogues are divided into two conversational floors, each involving two interlocutors: the left conversational floor consists of dialogue between the participant and the dialogue manager (DM), and the right consists of dialogue between the DM and the robot navigator (RN). The participant and RN never speak directly to or hear each other; instead, the DM acts as an intermediary passing communication between the participant and the RN. Of interest to our work, the left conversational floor (that which mimics our desired human-robot communication) is

[1]The SCOUT dataset and accompanying annotations presented here are available upon request.

	Left Conversational Floor		Right Conversational Floor	
#	**Participant**	**DM → Participant**	**DM → RN**	**RN**
1	move forward five feet			
2		I'm not sure if I can move forward five feet.		
3		I'll move as far forward as I can, ok?		
4	okay			
5		executing...		
6			Move forward as far as you can, up to 5 feet	
7				done
8		done		

Table 1: Navigation instruction initiated by the participant (#1), its clarification (#2-4), subsequent translation to a simplified form (Dialogue Manager (DM) to Robot Navigator (RN), #6), and acknowledgement of instructions (#5, 8) and execution by the RN (#7).

comprised of several potential modal expressions: "*move*" as an imperative; "*not sure"* as a negated epistemic; and "*can*" expressing a circumstantial ability.

All SCOUT speech data (collected from the participant and RN) are transcribed and time-aligned with text messages produced by the DM. SCOUT also includes annotations of *dialogue structure* (Traum et al., 2018) that allow for the characterization of distinct information states by way of sets of participants, participant roles, turn-taking and floor-holding, and other factors (Traum and Larsson, 2003). In total, SCOUT contains over 80 hours of human-robot dialogue from 83 participants.

2.3 Modal Expressions in Dialogue

As we are interested in modal meaning in context, we take a broad approach to the modal expressions we investigate, including modal verbs, attitude verbs, and imperatives. Most theories of modality in natural language take Kratzer (1981) as a starting point. Modal statements are interpreted relative to some *modal force*, e.g., necessity or various grades of possibility, and *conversational backgrounds*, e.g., realistic or normative. The traditional approach to attitude verbs treats them similarly to modals in a possible-worlds semantics (Hintikka, 1969): the verb specifies the set of accessible worlds (e.g., *believe* quantifies over worlds compatible with the beliefs of the attitude holder); quantification is taken to be universal.

As for imperatives, following Kaufmann (2019), imperatives should be treated similarly to modals; in fact, imperatives *are* modals. Any non-descriptive illocutionary force a modal proposition has comes from its context; the imperative modal operator presupposes that the context is non-descriptive. In contrast, Condoravdi and Lauer (2012) and Portner (2007) do not consider imperatives to be modals. Condoravdi and Lauer (2012) posit that each agent has an *effective preference structure* at any given world. Imperatives, then, are *public commitments* for the speaker's effective preference structure to be ordered in a certain way. Portner (2007), meanwhile, gives each interlocutor a *To-Do List*, a list of properties the agent is committed making true of themselves. The use of an imperative adds a property to the addressee's To-Do List. We adapt and motivate Portner's approach further in developing our formal theory in Section 5.

Previous work on modal expressions in dialogue is analogous to our own in prioritizing the discourse effect of such potentially ambiguous expressions, particularly those involving operators that take scope (Heim, 1982; Poesio, 1993; Lascarides and Asher, 2003). Authors concur that the semantic ambiguity of scopal operators (modals included) is typically reduced or absent in the context of human-human dialogue. Little work has focused on this resolution process in human-robot dialogue, instead focusing on documenting naturally occurring human language in this setting (Lukin et al., 2018b; Marge et al., 2020).

3 A Two-Level Annotation Scheme

The motivation for a two-level annotation scheme comes from the need to bridge formal semantic theories of modality and its interpretation with models that are actionable in the context of human-robot dialogue and adequately model the discourse. In this section we discuss the development of our annotation scheme, drawing on both fine-grained annotation of modality (Section 3.1) and the identification of speech acts specific to human-robot dialogue (Section 3.2). We present our final annotation scheme in Section 3.3.

Modal Value	Based on...	Example from SCOUT
Epistemic	...belief or knowledge.	I think you are more familiar with objects than I am. *(speaker = robot)*
Ability	...what someone or something can do.	Can you manipulate objects? *(speaker = human)*
Deontic	...what the rules, standards, or social norms state.	I need your help to decide which are important. <beep> *(speaker = robot)*
Bouletic	...someone's wishes or desires.	How far would you like me to turn left? *(speaker = robot)*
Teleological	...the achievement of goals.	I can move closer to take a picture. *(speaker = robot)*
Imperative	...syntactic form and imperative mood.	Please be aware of lag time. *(speaker = robot)*

Table 2: Modality type values for Level I annotation adapted from Rubinstein et al. (2013).

3.1 Level I: Fine-Grained Annotation of Modality

The first level of our annotation scheme is based on Rubinstein et al. (2013), who present a fine-grained annotation scheme of modal expressions and apply it to a subset of the MPQA corpus (Wiebe et al., 2005). The fine-grained nature of the annotation scheme results from the range of expressions the authors identify to carry modal meaning and the layers of information they annotate. We adapt the authors' understanding of modal expressions and their *Modality Type* category and accompanying values for our work, though we take into consideration the other elements they annotate.[2]

A modal expression is understood in this scheme as (i) an expression used to describe alternative ways the world could be, (ii) that has some sort of propositional argument (referred to as the prejacent), and (iii) is not associated with an overt attitude holder. Modality Type specifically categorizes the type of modality a modal expressions conveys in context. Seven fine-grained types are distinguished in Rubinstein et al.: *Epistemic*, *Circumstantial*, *Ability*, *Deontic*, *Bouletic*, *Teleological*, and *Bouletic/Teleological*. Before this classification is made, annotators first categorize each modal as belonging to one of two coarse-grained categories: *Priority* or *Non-Priority*. Priority picks out a conceptually motivated subclass of non-epistemic modalities: those that use some "priority" (a desire, a goal) to designate certain possibilities as better than others (Portner, 2009). For the MPQA corpus, annotators reliably agreed on only the highest level split between priority and non-priority interpretations (α=.89); Modality Type was quite challenging (α=.49).

The scheme we adapt for our Level I annotation is in Table 2. The modal expressions we target for annotation are broadly defined as any verb construction that conveys a modal meaning. Unlike the original scheme, we exclude modal nouns, adverbs, and adjectives and focus on verbs; we additionally annotate attitude verbs that have overt subjects (iii). This is both to provide coverage of different types of modal expressions we know to occur in our dialogue, as well as to simplify the annotation task, given the low annotator agreement of the original scheme. We additionally include the category *imperative* following work presented in Section 2.3, as a significant portion of our data is comprised of this type of utterance.

3.2 Level II: Speech Acts for Human-Robot Dialogue

The fine-grained annotation scheme developed by Rubinstein et al. (2013) is not sufficient for human-robot dialogue for two key reasons: (i) the scheme is geared towards modality in text, and thus does not consider how participant roles in spoken dialogue may impact modal meaning; and (ii) the shades of meaning the scheme pinpoint are not always meaningful in the context of achieving a specific task. Nevertheless, it is an ideal basis upon which to build a more complete understanding of modal interpretation in context.

The second level of our annotation thus encodes pragmatic information essential to successful interpretation of modal expressions in the context of dialogue. A robot first needs to understand if the illocutionary force of communications are (for example) commands, suggestions, or clarifications, which may not be obvious from the surface form of the human utterance alone. Furthermore, a robot needs to understand specific instructions such as how far to go and when, evaluate whether or not these instructions are feasible, and communicate and discuss the status of a given task in relation to a larger goal.

[2]These include *Environmental Polarity*, *Propositional Arguments*, *Source*, *Background*, *Modified Element*, *Degree Indicator*, *Outscoping Quantifier*, *Lemma*, and any additional notes from the annotator.

Type	Target	Scope	Level I: Value	Level II: Interpretation	Temporal Index
Modal expressions	lexical item with modal meaning	proposition	Modal types, Table 2	Speech acts, Table 8	local or global
Negation	not, n't, no	proposition	negation	NA	NA
Quantifiers	lexical quantifier	proposition	universal or existential	NA	NA

Table 3: An overview of our annotation scheme.

To this end, we incorporate the *speech act* inventory of Bonial et al. (2020) and *Dial-AMR*, a collection of 1122 utterances from the SCOUT corpus annotated with speech acts tailored to the robot in the search and navigation domain.[3] In delineating and defining their speech acts, the authors focus on the effects of an utterance relating to belief and obligation within human-robot dialogue (Traum, 1999; Poesio and Traum, 1998). Belief and obligation are not mutually exclusive, and utterances can and do often convey both the commitment to a belief and evoke an obligation in either the speaker or the hearer. These pragmatic effects are critical for agents navigating dialogue: in planning, agents can choose to pursue either goals or obligations and must reason about these notions so that the choice can be explained. Mutual beliefs about the feasibility of actions and the intention of particular agents to perform parts of that action are captured in the notion of *committed*, a social commitment to a state of affairs rather than an individual one (Traum, 1999). Incorporating notions of speaker intent into our annotation scheme is thus both practical and crucial to disambiguate the multiple meanings a modal expression can have.

There are fourteen possible values for the interpretation level of our annotation, all of which we preserve (though we expected and found not all to be compatible with modal expressions). The values, their relation to speaker and addressee commitments and obligations, and examples are given in Table 8 in Appendix A. These values map on to a set of 24 robot concepts, which designate the primitive concepts in the robot's knowledge ontology and include categories such as *ability*, *scene*, *environment*, *readiness*, and *help*.

3.3 Final Annotation Scheme

The goal of our final annotation scheme is to identify the range of naturally occurring modal expressions in task-based human-robot dialogue and to provide information about the use and interpretation of these expressions in context. In addition to modal expressions, we annotate negation and quantification for the purpose of detecting scope relations and meaning in dialogue more broadly in future work. Our approach acknowledges both the semantic richness of how modals are assigned interpretations in context (Rubinstein et al., 2013), as well as the situational grounding of the role an expression is playing in the task-oriented dialogue (Sarathy et al., 2019; Roque et al., 2020; Bonial et al., 2020). For this reason, we have developed a two-level annotation scheme that separates out the basic modal value of an expression from its eventual interpretation within a context.

We introduce a number of constraints to help pinpoint the interpretation of modal expressions in dialogue and to make annotation feasible for non-experts. First, we reduce the number of modality type values from Rubinstein et al. (2013) from seven to six, eliminating the *circumstantial* and combined *bouletic/teleological* values and adding a value for *imperative*. Our adaptation forces annotators to select a single, most salient category of modality type. The addition of an imperative value is due to the preponderance of this form in our data, and we discuss its broader implications in Section 5.

We compensate for the elimination of the circumstantial modal value by adding an additional layer of annotation: *temporal index*. The temporal index (TI) fixes the temporal reference of the modal expression based on the interaction of the modal with the semantics of the expressions it combines with (Condoravdi, 2001). In so doing, it designates how the expression of interest relates to the common ground between the speakers. There are two possible values for TI: (i) *Local* TI signifies that the utterance applies only to the immediate context; and (ii) *Global* TI signifies that the utterance adds meaningful, new information to the common ground that speakers should be aware of throughout the dialogue. A good diagnostic for this value is to ask how the subsequent response or action contributes to the understanding of the

[3]Dial-AMR augments standard Abstract Meaning Representation (AMR) (Banarescu et al., 2013) to map unconstrained language in natural human instructions to appropriate action specifications in the robot's limited repertoire.

Utterance	Level I: Value	Level II: Interpretation	Temporal Index
I **think** *you are more familiar with objects than I am.* (speaker = robot)	epistemic	assertion	global
Can *you manipulate objects?* (speaker = human)	ability	question	global
I **need** *your help to decide which objects are important.* (speaker = robot)	deontic	request	global
How far **would** *you* **like** *me to turn*? (speaker = robot)	bouletic	request	local
I **can** *move closer to take a picture.* (speaker = robot)	teleological	offer	local
Please ***be*** *aware of lag time* (speaker = robot)	imperative	request	global

Table 4: Example annotations with our annotation scheme for utterances for the six modality types we annotate, as defined in Table 2. Targets are in bold with scope in italics for each utterance.

utterance in context. For example, if a human commands "*move forward two feet*" to the robot, and the next action consists of the robot moving two feet forward, this is a *local* imperative (the task is completed and removed from the immediate context). Alternatively, if a human asks "*Robot, do you speak Arabic?*", both the question and answer to this provide lasting useful information: an intrinsic ability of the robot.

An overview of our final annotation scheme is seen in Table 3. A key question we aim to address with our scheme is the interaction of vagueness and ambiguity in natural language, or whether an utterance has one or many salient readings. The two primary levels of our annotation are comprised of linguistic categories well-known to be ambiguous: a modal expression can be both bouletic and teleological (*"I would like you to move forward so we can investigate the next room"*); while a speech act such as *"Why don't you ask for help?"* can be interpreted as a question and/or a suggestion. Similarly, TI introduces room for ambiguity: a human asking the robot "*Can you fit in that space?*" can be understood as both temporally local, in the sense that the robot moving into the space will clear this question from the immediate context; and global, in the sense that the subsequent response or action still contributes to the common ground as lasting information about the robot's size and abilities. Given the combined possibility for ambiguity in our annotation, we wanted to see whether or not clear interpretive distinctions emerge from the data. This information allows us to evaluate the ease with which future work integrating modality can be conducted (and whether or not is is a worthwhile endeavor to begin with). It also builds on the work of Bonial et al. (2020), whose scheme disfavors multiple possible interpretations that may nevertheless add important information to the dialogue.

4 Annotation Task

Our goal for the annotation task was two-fold: (i) to provide coverage of the data to quantitatively assess the kind and frequency with which modal expressions are used and interpreted by speaker type; and (ii) to qualitatively assess instances where modal usage is unexpected. This second goal is situated within a larger goal of understanding and automating the interpretation of scope in human-robot dialogue.

Four annotators were trained to apply the annotation scheme following the annotation guidelines and with two example annotated transcripts. Each annotator annotated 70 experimental transcripts, of which 16 transcripts overlapped with one of the other three annotators. In total, 248 transcripts were annotated: 32 by two annotators, and the remaining 216 by a single annotator. Annotators were instructed to only annotate the left conversation floor (Table 1), as this is designed to mimic automated human-robot dialogue. For each category of annotation except *scope*, annotators were provided with a drop-down menu that allowed them to easily restrict their choice of value; scope was manually annotated. For utterances that contained multiple types, annotators were instructed to annotate each type separately. There were a total of 48,168 utterances in the left conversation floor (22,259 human, and 25,909 robot) across all transcripts, for an average of 194.23 utterances per transcript. Examples of final annotations are given in Table 4.

To evaluate our annotation scheme, we calculated a number of inter-annotator agreement metrics on the 32 transcripts annotated by two annotators. First, we calculated the proportion of annotations for which the annotators agreed on the target. Among those annotations where the annotators agreed on the target, for each pair of annotators, we calculated the string overlap[4] between the scopes identified by each

[4]We define string overlap between two strings s and t to be $\frac{\text{lcs}(s,t)}{\text{maxlen}(s,t)}$, where $\text{lcs}(s,t)$ is the longest common substring in s and t, and $\text{maxlen}(s,t)$ is the length of the longer of s and t.

Category	Median κ (range)
Type	1.0000 (0.9909-1.0000)
Value	0.9458 (0.7683-1.0000)
Interpretation	0.8493 (0.6554-1.0000)
Temporal index	0.8612 (0.7409-0.9616)

Table 5: Cohen's kappa for each category, for each annotator pair.

Type	Count
Modal expressions	18,125
· *Attitude verbs*	551
· *Imperatives*	13,440
· *Other modal verbs*	4,134
Negation	853
Quantifiers	478

Table 6: Distribution of modal expressions, negation, and quantifiers in our data.

annotator, and Cohen's kappa (Cohen, 1960) for type, value, interpretation, and temporal index.

4.1 Results

Of the 3,959 annotations in the 32 shared transcripts, annotators agreed on the target for 3,470 (87.65%). Among each pair of annotators, the string overlap for scope ranged from 83.31% to 91.59% (median 85.97%). Table 5 describes Cohen's kappa (median and range) (IAA) for each category.

After calculating IAA, we adjudicated the shared transcripts and combined them with our singly-annotated transcripts to form a gold standard. In total, 18,073 utterances (37.52%) contained one or more annotations. There were 19,456 total annotations, for an average of 1.08 annotations per annotated utterance. The distribution of modal expressions (including attitude verbs, imperatives, and other modal verbs), negation, and quantifiers is shown in Table 6. Additional result tables, describing the classification of modal expressions by speaker, value, interpretation, and temporal index, are presented in Appendix B.

4.2 Discussion

Several data points are of immediate interest from our annotation results. First, there are several asymmetries in how humans and 'robots' employ modality and illocutionary force. Humans use many more imperative modal forms than robots (13,257/120), 59.56% of their total utterances (Table 9); this finding correlates with humans using more command speech acts than robots (13,616/121) and more command speech acts than any other speech act type (Table 10), confirming findings from Marge et al. (2017). In contrast, the SCOUT robot employs teleological (1,591/432) and bouletic (184/19) modal values more frequently than the human; these tend to be in the form of making offers to perform certain actions (bouletic, Table 11) or assertions, promises, questions, and requests related to the task goals (teleological). For speech acts overall, the robot most commonly employs assertions (1,167) and promises (1,001).

Overall, our IAA scores are higher than we expected (Table 5). Though this is likely due in large part to the repetitive nature of the SCOUT data, it both validates our annotation scheme for future use and sheds light on the attested interaction of modal expressions and their interpretation. As expected, ability modals demonstrate the most flexibility in use in our data: they are employed for eight of the fourteen speech act values found. Teleological modals are also quite flexible: these are employed for ten of the fourteen speech act values (though only in single instances for two values). Epistemic modals pattern to either assertions or questions, while bouletic modals primarily comprise offers. With regards to TI, the majority of utterances are local and relevant to the immediate context rather than adding lasting information to the common ground; this imbalance is less pronounced, however, for ability and epistemic modals.

Other phenomena of interest from our data involve modal operators and their scope. For example, there were 298 utterances containing both a modal expression and a negation. Of those, a negation scopes over a modal in 227 ("*I couldn't hear everything you said*"), while in 53, a modal scopes over a negation ("*Can you first scan the area you haven't scanned yet?*"). We note that anaphora and coreference on one hand ("*Do that again*", "*That sounds good*"), and implicit arguments on the other ("*Repeat ∅*", "*Yes I would ∅*"), are quite challenging with regards to identifying the proposition in the scope of the modal operator. In contrast, we also find utterances where only the proposition is explicit, and the operator implicit ("*45 degrees*", "*Picture*") These phenomena fall under the umbrella of underspecification, an enduring challenge of creating meaningful natural language representation that must nevertheless be actionable in settings like HRI. Finally, corrected or disjoint scope ("*Can you turn 90 degrees left... I mean right*") and coordination ("*Can you go back inside and take a picture*") also pose challenges to scope in

dialogue, especially in the context of sentence-based meaning representation (Pustejovsky et al., 2019).

Finally, we note some utterances that our annotation scheme alone cannot account for. These include conditional utterances such as "*If you can turn around and take a photo so I can have a clear picture*" interpreted as commands. We note for now that these utterances exemplify our ambiguity challenge: the modal *can* has both ability and teleological meanings, while the utterance can function as both a request (given its conditional nature) and a command (given that it is uttered by the human).

5 Towards a Formal Theory of Modality in Human-Robot Dialogue

Here, we sketch the beginnings of a formal theory of modality in human-robot dialogue built upon our annotation findings in Section 4. As stated before, this work takes a step towards automating the interpretation of frequently ambiguous expressions in context and mapping this interpretation to actionable representations in the human-robot context.

5.1 Desiderata

In typical human dialogue, there is a shared understanding of both an utterance meaning (content) and the speaker's meaning in the specific context (intent). This is what our annotation has captured. The ability to link these two dynamically is the act of situationally grounding meaning to the local context, or *establishing the common ground* between interlocutors (Stalnaker, 2002; Asher and Gillies, 2003; Tomasello and Carpenter, 2007). The common ground represents the mutual knowledge, beliefs, and assumptions of the participants that result from co-situatedness, co-perception, and co-intent. Robust human-robot dialogue requires a unique process of alignment to facilitate human-like interaction, including the recognition and generation of expressions through multiple modalities (language, gesture, vision, action); and the encoding of *situated meaning* (Dobnik et al., 2013; Pustejovsky et al., 2017; Krishnaswamy et al., 2017; Hunter et al., 2018). Specifically, this entails outlining three key aspects of common ground interpretation: (i) the situated *grounding* of expressions in context; (ii) an interpretation of the expression contextualized to the *dynamics* of the discourse; and (iii) an appreciation of the *actions and consequences* associated with objects in the environment. Here, we address (ii) first, before moving on to (i) and (iii).

5.2 Dynamic Interpretation of Modal Expressions

An account of how modal expressions are used in discourse needs to capture their command-force "context change potential" (CCP), usually modeled as a function from input contexts to output contexts, as well as how this relates to an agent behaving rationally and cooperatively relative to their commitments (Section 3.2). An adequate model of the common ground in human-robot dialogue will especially require a satisfactory account of imperatives, as these are so frequent and directly impact goal achievement.

Here, we follow Portner (2007) in the idea that imperatives technically do not add to the common ground (and are technically not modals), while modals do (as they can be evaluated as true or false). Imperatives are instead evaluated relative to the addressee's *To-Do list* (TDL), a list of properties (not propositions). TDL is nevertheless a contextual resource for the interpretation of priority modals, analogous to the common ground for epistemic modals. An imperative specifically adds an addressee-restricted property to a hearer's TDL such that the hearer should act so as to make as many items on TDL *true as feasible*. This is based on a mutual assumption between the participants that each will try to bring it about that they have each of these properties. For example, if a given property corresponds to an action ([$\lambda w \lambda x.x$ moves forward two feet in w]), the TDL represents the actions that an agent α is committed to taking. The TDL function T assigns to each α in the conversation a set of properties $T(\alpha)$. The canonical discourse function of an imperative clause ϕ_{imp} is then to add $[\![\phi_{\text{imp}}]\!]$ to $T(addressee)$, where C is a context of the form $\langle CG, Q, T\rangle$: $C + \phi_{\text{imp}} = \langle CG, Q, T[addressee/(T(addressee) \cup \{[\![\phi_{\text{imp}}]\!]\})]\rangle$. More details are in Appendix C.

In other words, imperatives make reference to an additional component of the context set: the TDL, formalized by $T(\alpha)$. TDLs are structured with different "flavors" similar to how ordering sources differ for modals. Thus, each participant in a conversation possesses multiple TDLs that correspond to priority types: a teleological TDL represents goals; a bouletic TDL, desires; and a deontic TDL, obligations. In addition to assuming these, we propose another flavor of TDL specific to human-robot dialogue: a shared TDL that represents shared goals, desires, and obligations. Both individual and shared TDLs in

		Modal (I)	Interpretation (II)	TI	Logical representation and interpretive process
a.	Can $\overbrace{\text{you repeat that}}^{p}$? *(speaker = robot)*	ability	request	local	**interpretation**: ability inquiry to human; p is known in CG; ϕ ::= **can**(p) ϕ offered for consideration in $T_{\text{local-human}}$ = $\lambda w \lambda x . x$ repeats utterance in w
b.	Can $\overbrace{\text{you send a picture}}^{p}$? *(speaker = human)*	ability	command	local	**interpretation**: ability inquiry evaluated as potential command; p is stated in CG; ϕ ::= **can**(p) ϕ interpreted as imperative in $T_{\text{local-robot}}$; represents action $\lambda w \lambda x . x$ sends picture in w
c.	$\overbrace{\text{You}}$ $\overbrace{\text{can}}$ $\overbrace{\text{tell me to}}^{p}$ $\overbrace{\text{move to a certain object.}}$ *(speaker = robot)*	ability	suggestion	global	**interpretation**: ϕ ::= **can**(p) ϕ added to $T_{\text{global-shared}}$
d	Do $\overbrace{\text{you speak Arabic}}^{p}$? *(speaker = human)*	ability	question	global	**interpretation**: ϕ ::= **can**(p) ϕ does not update TDL

Table 7: Interpretive variation of modal type *ability* in relation to speaker and temporal index (TI) with corresponding mappings to logical representation in our proposed context set $\langle CG, Q, T \rangle$.

our scheme ought to possess local and global temporal indices, reflecting our annotation and the discrete and continuous planning functions of robots they correspond to (Chai et al., 2018).

These intuitions can be formalized in the interpretations in Table 7. We use ability modals as an example, as they demonstrate a range of flexibility in their illocutionary force in our data. For present purposes, we understand the denotation of the modal auxiliary *can* as: $[\![can]\!] = \lambda p \exists w(w \in \textit{MB}(e) : q(w))$, where MB (*modal base*) represents the set of states that are compatible with the utterance (Hacquard and Cournane, 2016). For example, the temporal indices in of *local* (a,b) and *global* (c,d) force circumstantial and epistemic interpretations, respectively. The additional interpretations in Table 7 fall out from our context set $\langle CG, Q, T \rangle$. The context set includes vital information such as the speaker/addressee relationship particular to the human-robot context, in which the human is endowed with more authority; the question or goal under discussion (Ginzburg, 1995); and other properties of the common ground, described in 5.3. The reverse mapping can be formalized, too: ability, bouletic, deontic, and teleological modals can all map on to request speech acts (Table 8), though their logical representations will differ. As Bonial et al. (2020) use AMR to represent their speech act inventory (footnote 3), we plan to extend our work by translating AMR into first-order-logic for simpler mapping to robot action (Lai et al., 2020). Logical differences between modal categories will then be captured in our FOL translations and assist the robot in understanding the mappings between modal expressions and speech acts.

5.3 Situated Grounding and Modal Meaning

As noted in Section 2.2, dialogues in SCOUT were collected to mimic the setting of a low-bandwidth reconnaissance or search-and-navigation operation. A participant verbally instructs a robot at a remote location, guiding the robot to explore a physical space. The sensors and video camera on-board the robot populate a map as it moves, enabling it to describe that environment and send photos at the participant's request, but the communications bandwidth prohibits real-time video streaming or direct tele-operation. The robot is assumed capable of performing low to intermediate level tasks, but not more complex tasks involving multiple or quantified goals without clear instruction. The experiment used a Clearpath Robotics Jackal, fitted with an RGB camera and LIDAR sensors, to operate in the environment (Marge et al., 2017).

Given this as background, we assume that both robot and human are aware of these capabilities and that they are in the common ground, entering into the dialogues under discussion. From the robot's perspective, the objects in the environment present opportunities for interaction, exploration, and manipulation. These are modally contingent actions that a situation presents to an agent by virtue of the objects it encounters. The contextual meaning for many modal expressions will be interpreted relative to such object knowledge.

For these reasons, it is useful to think of objects as providing *habitats*, which are situational contexts

or environments conditioning the object's *affordances*, which may be either "Gibsonian" affordances (Gibson et al., 1982) or "Telic" affordances (Pustejovsky, 1995). A habitat specifies how an object typically occupies a space (Pustejovsky, 2013). Affordances are used as attached behaviors, which the object either facilitates by its geometry (Gibsonian) or purposes for which it is intended to be used (Telic). For example, a Gibsonian affordance for [[CUP]] is "grasp," while its Telic affordance is "drink from." Similarly, in SCOUT's environment, a "doorway" affords passage to another room, unless it is blocked by an object or closed. Hence, when asked: "*Can you go through the doorway?*", the modal force is taken as a query over its situational (or local) ability, given what the speaker already knows about the robot's navigation capabilities. An example representation of the affordances of a "doorway" is given in Appendix D. In a similar manner, the question: "*Do you speak Arabic?*" is interpreted as a general ability modal, motivated by the situational awareness of Arabic script identified in the picture the robot sent. That is, linguistic signs afford decoding or interpretation, which prompts the modal reference to the ability to speak the language associated with the affording script (Sundar et al., 2010; Krippendorff, 2012).

5.4 Putting it All Together

We have sketched components that allow us to conceptualize how to formalize the key aspects of common ground interpretation we outlined in 5.1. A proper treatment of modal expressions in human-robot dialogue will integrate both the dynamic semantics of 5.2 as well as how the grounding of objects explained in 5.3 impacts this interpretation by allowing the robot to reason about abilities, actions, and consequences.

Nevertheless, work remains. The data we present support findings that humans tend towards a less verbose style of communication with robots than with other humans (Lukin et al., 2018b); and that humans spend less time updating beliefs and planning with robots than with other humans (Marge et al., 2020). In contrast, the surrogate 'robot' of our data orients its utterances towards goal-completion and general cooperation, behaving in a more constructive and polite manner. If we expect future robots to learn behavior and language use through interaction, these results are problematic. This paradox suggests that other avenues for the learning of modal expressions ought to be explored, specifically those that leverage existing semantic representations and modal ontologies such as ours to endow the robot with semantic knowledge prior to interaction. From a practical standpoint, modal expressions allow a robot to determine the meaning of a natural language utterance, generate a goal representation with reference to existing goals, and produce an action sequence to achieve the new goal if possible (Dzifcak et al., 2009). From a social standpoint, modal expressions reflect and create participant relations, impacting factors such as trust and openness that indirectly foster successful collaboration (Lukin et al., 2018b; Lucas et al., 2018). Thus, the work we present here is very much worth exploring further.

6 Conclusion

In this paper, we present a two-level annotation scheme for modality as used in situated human-robot dialogues relating to search and navigation. Our annotation scheme captures both the semantic content of modal expressions as well as their pragmatic function relevant to speaker intent in discourse. Results from our annotation task demonstrate that our annotation scheme is valid and expressive, as well as both practical and transparent; it also gives us novel insight into the interaction between modality and illocutionary force in our setting. Our work can be extended to future, automated pipelines for human-robot dialogue that incorporate modal expressions within a formal common ground.

Acknowledgements

We are grateful to Clara Wan Ching Ho, Bailey Johnson, David Meier, and Christoph Otto for their valiant annotation efforts. Linxuan Yang was instrumental in providing input throughout the process of developing the annotation scheme, as well.

References

Nicholas Asher and Anthony Gillies. 2003. Common ground, corrections, and coordination. *Argumentation*, 17(4):481–512.

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin, and Scott Miller. 2012. Modality and negation in SIMT use of modality and negation in semantically-informed syntactic MT. *Computational Linguistics*, 38(2):411–438.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-amr: Abstract meaning representation for dialogue. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 684–695.

Joyce Y. Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cleo Condoravdi and Sven Lauer. 2012. Imperatives: Meaning and illocutionary force. *Empirical issues in syntax and semantics*, 9:37–58.

Cleo Condoravdi. 2001. Temporal interpretation of modals-modals for the present and for the past. In *The construction of meaning*. Citeseer.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2013. Modelling language, action, and perception in type theory with records. In *Constraint Solving and Language Processing*, pages 70–91. Springer.

J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn. 2009. What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In *2009 IEEE International Conference on Robotics and Automation*, pages 4163–4168.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of the fourteenth conference on computational natural language learning–Shared task*, pages 1–12.

James Jerome Gibson, Edward S Reed, and Rebecca Jones. 1982. *Reasons for realism: Selected essays of James J. Gibson*. Lawrence Erlbaum Assoc Incorporated.

Jonathan Ginzburg. 1995. Resolving questions, i. *Linguistics and philosophy*, 18(5):459–527.

Valentine Hacquard and Ailis Cournane. 2016. Themes and variations in the expression of modality. In *Proceedings of NELS*, volume 46, pages 21–42.

Irene Heim. 1982. *The semantics of definite and indefinite noun phrases*. Ph.D. thesis, University of Massachusetts Amherst, Department of Linguistics.

Jaakko Hintikka. 1969. Semantics for propositional attitudes. In *Models for modalities*, pages 87–111. Springer.

Julie Hunter, Nicholas Asher, and Alex Lascarides. 2018. A formal semantics for situated conversation. *Semantics and Pragmatics*, 11.

Magdalena Kaufmann. 2019. Fine-tuning natural language imperatives. *Journal of Logic and Computation*, 29(3):321–348.

Angelika Kratzer. 1981. The notional category of modality. *Words, Worlds, and Contexts: New Approaches in Word Semantics*, 6:38.

K. Krippendorff. 2012. Discourse and the materiality of its artifacts. In *Matters of communication: Political, cultural, and technological challenges to communication theorizing*, pages 23–46. Hampton Press.

Nikhil Krishnaswamy, Pradyumna Narayana, Isaac Wang, Kyeongmin Rim, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Jaime Ruiz, Ross Beveridge, Bruce Draper, and James Pustejovsky. 2017. Communicating and acting: Understanding gesture in simulation semantics. In *12th International Workshop on Computational Semantics*.

Kenneth Lai, Lucia Donatelli, and James Pustejovsky. 2020. A continuation semantics for abstract meaning representation. In *Proceedings of the Second International Workshop on Designing Meaning Representations*.

Alex Lascarides and Nicholas Asher. 2003. Imperatives in dialogue. *Pragmatics and Beyond New Series*, pages 1–24.

Changsong Liu and Joyce Y. Chai. 2015. Learning to mediate perceptual differences in situated human-robot dialogue. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Gale M. Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2018. Getting to know each other: The role of social dialogue in recovery from errors in social robots. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction*, pages 344–351.

Stephanie Lukin, Felix Gervits, Cory Hayes, Pooja Moolchandani, Anton Leuski, John G. Rogers III, Carlos Sanchez Amaro, Matthew Marge, Clare Voss, and David Traum. 2018a. ScoutBot: A dialogue system for collaborative navigation. In *Proceedings of ACL 2018, System Demonstrations*, pages 93–98, Melbourne, Australia, July. Association for Computational Linguistics.

Stephanie Lukin, Kimberly Pollard, Claire Bonial, Matthew Marge, Cassidy Henry, Ron Artstein, David Traum, and Clare Voss. 2018b. Consequences and factors of stylistic differences in human-robot dialogue. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 110–118, Melbourne, Australia, July. Association for Computational Linguistics.

Ana Marasović and Anette Frank. 2016. Multilingual modal sense classification using a convolutional neural network. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 111–120, Berlin, Germany, August. Association for Computational Linguistics.

Matthew Marge, Claire Bonial, Brendan Byrne, Taylor Cassidy, A. William Evans, Susan G. Hill, and Clare Voss. 2016. Applying the Wizard-of-Oz technique to multimodal human-robot dialogue. In *RO-MAN 2016: IEEE International Symposium on Robot and Human Interactive Communication*.

Matthew Marge, Claire Bonial, Ashley Foots, Cory Hayes, Cassidy Henry, Kimberly Pollard, Ron Artstein, Clare Voss, and David Traum. 2017. Exploring variation of natural human commands to a robot in a collaborative navigation task. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 58–66, Vancouver, Canada, August. Association for Computational Linguistics.

Matthew Marge, Felix Gervits, Gordon Briggs, Matthias Scheutz, and Antonio Roque. 2020. Let's do that first! a comparative analysis of instruction-giving in human-human and human-robot situated dialogue. In *The 24th Workshop on the Semantics and Pragmatics of Dialogue*. Brandeis University.

Roser Morante and Walter Daelemans. 2012. Annotating modality and negation for a machine reading evaluation. In *CLEF (Online Working Notes/Labs/Workshop)*, pages 17–20.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational linguistics*, 38(2):223–260.

Massimo Poesio and David Traum. 1998. Towards an axiomatization of dialogue acts. In *Proceedings of the Twente Workshop on the Formal Semantics and Pragmatics of Dialogues (13th Twente Workshop on Language Technology*. Citeseer.

Massimo Poesio. 1993. *Assigning a Scope to Operators in Dialogues*. Ph.D. thesis, University of Rochester, Department of Computer Science.

Paul Portner. 2007. Imperatives and modals. *Natural language semantics*, 15(4):351–383.

Paul Portner. 2009. *Modality*, volume 1. Oxford University Press.

Vinodkumar Prabhakaran, Michael Bloodgood, Mona Diab, Bonnie Dorr, Lori Levin, Christine D. Piatko, Owen Rambow, and Benjamin Van Durme. 2012. Statistical modality tagging from rule-based annotations and crowdsourcing. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 57–64, Jeju, Republic of Korea, July. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, et al. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91.

James Pustejovsky, Nikhil Krishnaswamy, Bruce Draper, Pradyumna Narayana, and Rahul Bangar. 2017. Creating common ground through multimodal simulations. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

James Pustejovsky, Nianwen Xue, and Kenneth Lai. 2019. Modeling quantification and scope in abstract meaning representations. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 28–33.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

James Pustejovsky. 2013. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10. ACL.

Laurel Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, 1(1).

Antonio Roque, Alexander Tsuetaki, Vasanth Sarathy, and Matthias Scheutz. 2020. Developing a corpus of indirect speech act schemas. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 220–228.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Dan Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. In *Proceedings of the IWCS 2013 workshop on annotation of modal meanings in natural language (WAMM)*, pages 38–46.

Vasanth Sarathy, Thomas Arnold, and Matthias Scheutz. 2019. When exceptions are the norm: Exploring the role of consent in hri. *ACM Transactions on Human-Robot Interaction (THRI)*, 8(3):1–21.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227.

Roser Saurı, Marc Verhagen, and James Pustejovsky. 2006. Annotating and recognizing event modality in text. In *Proceedings of 19th International FLAIRS Conference*.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5-6):701–721.

Qi Su, Chu-Ren Huang, and Helen Kaiyun Chen. 2010. Evidentiality for text trustworthiness detection. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 10–17.

S. Shyam Sundar, Qian Xu, and Saraswathi Bellur. 2010. Designing interactivity in media interfaces: A communications perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2247–2256.

Michael Tomasello and Malinda Carpenter. 2007. Shared intentionality. *Developmental science*, 10(1):121–125.

David Traum and Staffan Larsson. 2003. The information state approach to dialogue management. In Jan van Kuppevelt and Ronnie W. Smith, editors, *Current and new directions in discourse and dialogue*, pages 325–353. Springer.

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory Hayes, and Susan Hill. 2018. Dialogue structure annotation for multi-floor interaction. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 104–111, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

David Traum. 1999. Speech acts for dialogue agents. In Anand Rao and Michael Wooldridge, editors, *Foundations of Rational Agency*, pages 169–201. Kluwer.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

A Speech Acts for Human-Robot Dialogue

Interpretation	Commitments & Obligations	Examples
1. Question	Speaker (S) committed to desire to know answer; Addressee (A) obliged to respond to question	*Do you see foreign writing?* (ability) *Are you able to move that orange cone in front of you?* (ability)
2. Assertion	S committed to a state of affairs	*I think you are more familiar with shoes than I am.* (epistemic) *I'm not able to manipulate objects.* (ability)
3. Offer	S committed to feasibility of plan of action; A obliged to consider action and respond	*Would you like me to take a picture?* (bouletic) *I will move forward one foot, ok?* (teleological)
4. Promise	S committed to feasibility of plan of action and obliged to do action	*I will send a picture.* (teleological) *I'll give you feedback on what I'm doing.* (teleological)
5. Command	S committed to desire for A to do something and feasibility of action; A obliged to do action	*Back up three feet.* (imperative) *Help!* (imperative)
6. Open-Option	S committed to feasibility of action(s)	*If you describe an object, you can help me learn what it is.* (ability) *You can tell me to move a certain distance or to move to an object.* (ability)
7. Request	S committed to desire for A to do something and feasibility of action; A is obliged to consider action and respond	*Can you repeat that?* (ability) *I need your help to find shoes.* (deontic)
8. Accept/Reject	S committed to a state of general acceptance or rejection	*I think you are correct.* (epistemic) *Yes please proceed.* (imperative)
9. Greeting	S committed to recognizing presence of A and willingness to interact	*Hello!*
10. Gratitude	S committed to state of gratitude	*Thanks, teammate!*
11. Regret	S committed to state of regret	*Sorry robot I meant west.* (epistemic)
12. Judgment	S committed to evaluative stance	*The containers look like something that could be moved.* (ability) *Possible equipment or storage space they might need for something for gathering.* (teleological)
13. Mistake	S committed to acknowledging error	*I mean twenty degrees to the left.* (teleological) *Actually i meant turn left fifty degrees.* (teleological)
14. Hold Floor	S committed to holding conversational floor for continued speech	(not found in our data)

Table 8: Speech Act Lexicon from Bonial et al. (2020), adapted here for Level II of our annotation. Examples are from the SCOUT corpus with modal values in parentheses when applicable. Note: A response (*Request*) might be by doing the action, rejecting it, accepting it, or discussing desirability. Expressive types (*Request* and subsequent rows) are left unspecified as to the resulting obligations and some further commitments, since some derive as much from context and committed mental state as well as the act itself, and some are culture-specific. For example, an acceptance of a Request generally commits the accepter to act, and an acceptance of an Offer generally commits the offerer to act.

B Additional Annotation Results

The tables below describe the classification of modal expressions by speaker (Tables 9, 10, and 12), value (Tables 9, 11, and 13), interpretation (Tables 10, 11, and 14), and temporal index (Tables 12, 13, and 14).

Value	Speaker	
	Human	Robot
Ability	743	1,103
Bouletic	19	184
Deontic	7	6
Epistemic	287	376
Imperative	13,257	120
Teleological	432	1,591

Table 9: Modal expressions, by speaker and value.

Interpretation	Speaker	
	Human	Robot
Accept/Reject	5	0
Assertion	381	1,167
Command	13,616	121
Gratitude	1	0
Greeting	0	0
Hold Floor	0	0
Judgment	5	1
Mistake	7	0
Offer	13	250
Open-Option	4	157
Promise	57	1,001
Question	152	453
Regret	1	0
Request	503	230

Table 10: Modal expressions, by speaker and interpretation.

Interpretation	Value					
	Ability	Bouletic	Deontic	Epistemic	Imperative	Teleological
Accept/Reject	0	0	0	1	4	0
Assertion	404	16	12	609	1	506
Command	330	7	0	1	13,326	73
Gratitude	0	0	0	0	1	0
Greeting	0	0	0	0	0	0
Hold Floor	0	0	0	0	0	0
Judgment	3	0	0	2	0	1
Mistake	0	1	0	2	1	3
Offer	77	143	0	0	0	43
Open-Option	124	4	0	0	0	33
Promise	527	0	1	0	41	489
Question	113	30	0	47	0	415
Regret	0	0	0	0	0	1
Request	268	2	0	1	3	459

Table 11: Modal expressions, by value and interpretation.

Speaker	Temporal Index	
	Global/common	Local/utterance
Human	166	14,579
Robot	581	2,799

Table 12: Modal expressions, by speaker and temporal index.

Value	Temporal Index	
	Global/common	Local/utterance
Ability	330	1,516
Bouletic	5	198
Deontic	6	7
Epistemic	281	382
Imperative	47	13,330
Teleological	78	1,945

Table 13: Modal expressions, by value and temporal index.

Interpretation	Temporal Index	
	Global/common	Local/utterance
Accept/Reject	0	5
Assertion	448	1,100
Command	49	13,688
Gratitude	0	1
Greeting	0	0
Hold Floor	0	0
Judgment	1	5
Mistake	0	7
Offer	43	220
Open-Option	131	30
Promise	36	1,022
Question	29	576
Regret	0	1
Request	10	723

Table 14: Modal expressions, by interpretation and temporal index.

C Further To-Do List Semantics (Portner, 2007)

1. Requirement constraint (modal base):
   $\forall w[(w \in \cap CG \wedge \neg\exists[w' \in \cap CG \wedge w' <_i w]) \rightarrow w \in [[S]]$

2. Ordering source to determine robot course of action (To-Do List):
   $T(r) = \{C(x,h)_i, M(x,f)_k\}$, where C is a cooperate relation with a temporal index i and M is a move event with a temporal index of k

3. *Pragmatic function of imperatives*

   (a) The To-Do List function T assigns to each participant α in the conversation a set of properties $T(\alpha)$.

   (b) The canonical discourse function of an imperative clause ϕ_{imp} is to add $[\![\phi_{\text{imp}}]\!]$ to $T(addressee)$. Where C is a context of the form $\langle CG, Q, T\rangle$:
   $C + \phi_{\text{imp}} = \langle CG, Q, T[addressee/(T(addressee) \cup \{[\![\phi_{\text{imp}}]\!]\})]\rangle$

4. Partial ordering of worlds by TDL compatible with CG ($\in \bigcap$ CG = context set):
   For any $w_1, w_2 \in \bigcap$ CG and any participant i, $w_1 < w_2$
   iff for some $P \in T(i)$, $P(w_2)(i) = 1$ and $P(w_1)(i) = 0$,
   and for all $Q \in T(i)$, if $Q(w_1)(i) = 1$, then $Q(w_2)(i) = 1$

5. Agent's commitment:
   For any participant i, the participants in the conversation mutually agree to deem i's actions rational and cooperative to the extent that those actions in any world $w_1 \in \bigcap$ CG such that $w_1 <_i w_2$

D Example Representation for Dialogue Concept "Doorway"

(1)
$$
\left[\begin{array}{l}
\textbf{doorway} \\
\text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{doorway} \\ \text{TYPE} = \textbf{aperture}\end{array}\right] \\
\text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{rectangular\_prism} \\ \text{COMPONENTS} = \textbf{nil} \\ \text{CONCAVITY} = \textbf{flat} \\ \text{ROTATSYM} = \{X, Y, Z\} \\ \text{REFLECTSYM} = \{XY, XZ, YZ\}\end{array}\right] \\
\text{HABITAT} = \left[\begin{array}{l}\text{INTR} = \left[\begin{array}{l}\text{UP} = align(Y, \mathcal{E}_Y) \\ \text{FRONT} = open\end{array}\right] \\ \text{EXTR} = \ldots\end{array}\right] \\
\text{AFFORD\_STR} = \left[\begin{array}{l}\text{A}_1 = H_1 \rightarrow [walk\_through]R\end{array}\right] \\
\text{EMBODIMENT} = \left[\begin{array}{l}\text{SCALE} = >\textbf{agent} \\ \text{MOVABLE} = \textbf{false}\end{array}\right]
\end{array}\right]
$$